NLP 모델 기반 뉴스 요약봇 제작

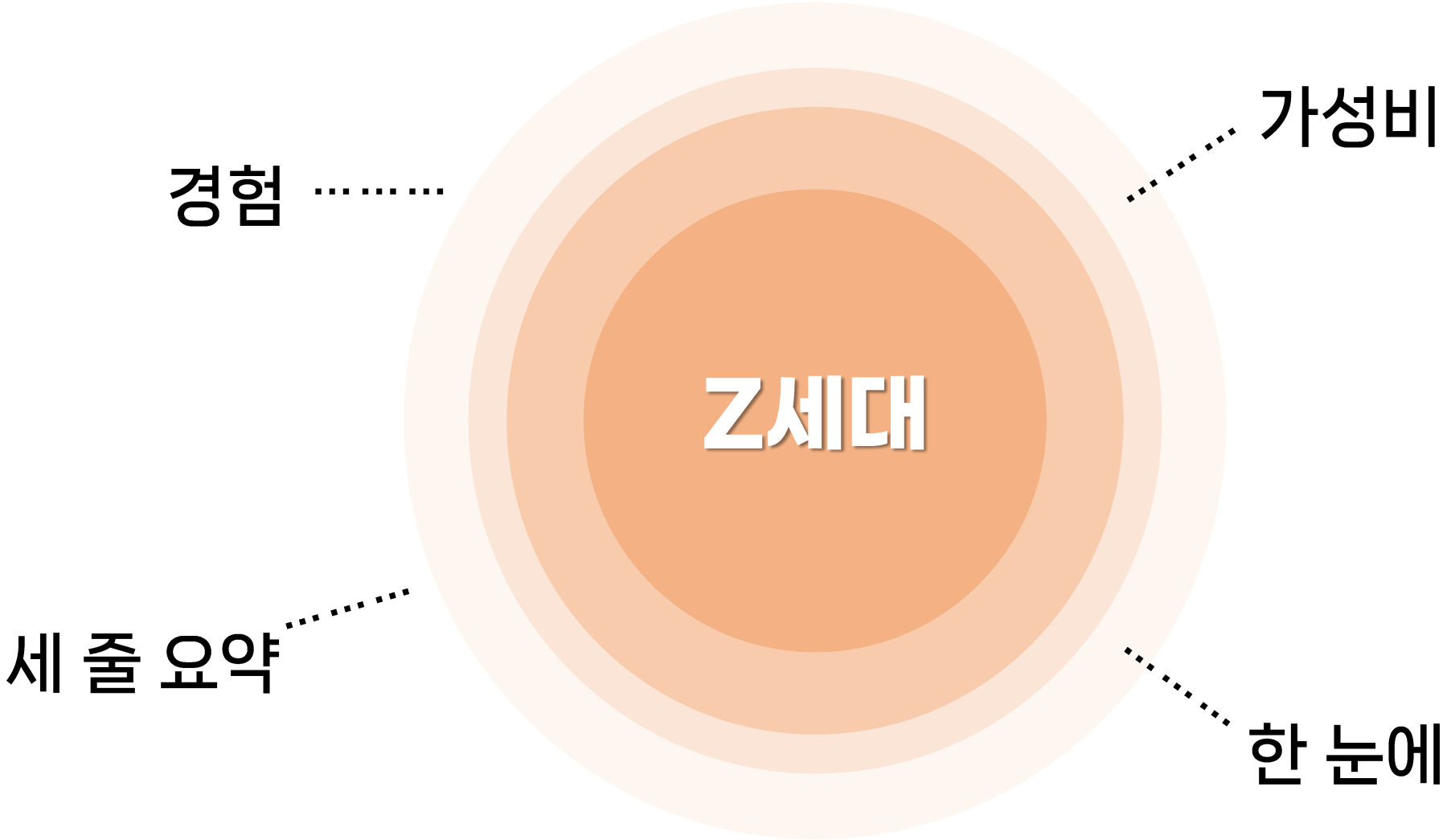
3조 곽형민 김재현 박성혜 이창재 전민정

alpaco

CONTENTS

1. 배경

1020 성향



Z세대, 숏폼 컨텐츠 성행

- 연령대가 낮을수록 10분 미만 숏폼 동영상 선호
- 1020세대의 동영상 선호 시청 길이는 15분 내외
- Z세대 80% 평일 75분 숏폼 콘텐츠 시청
- 네이버, 카카오 등 국내 기업 숏폼 콘텐츠 서비스 강화

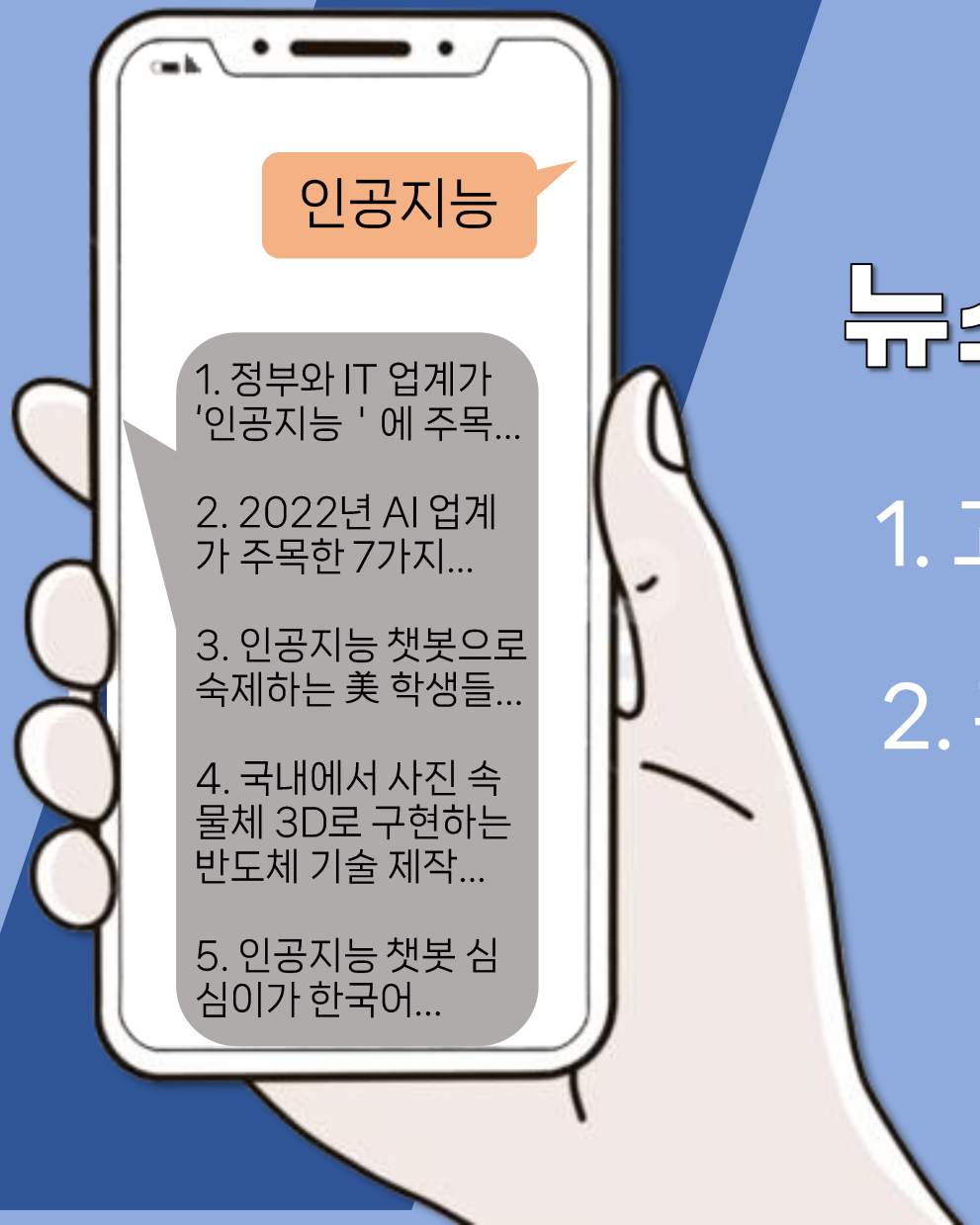
뉴스 영역까지 등장한 숏폼 컨텐츠

- 네이버 뉴스에 새로 등장한 20대 뉴스판
- 키워드는 한눈에, 이슈 파악은 짧고 쉽게
- 각종 언론사에서 숏폼 컨텐츠 제작에 촉각
- 유튜브, 틱톡, 네이버 뉴스에 숏폼 컨텐츠 선보임

한 눈에

“더욱 [간단하게] 만들자”

빠르게



뉴스 요약봇

1. 고객이 관심 키워드 검색하면

2. 관련 뉴스 최신 top10 요약

2. 제작 과정

Train 데이터

- 출처 : AI HUB

- 소개

데이터 영역	한국어	데이터 유형	텍스트
데이터 형식	txt	데이터 출처	신문, 보도자료, 간행물, 문학, 연설문 등
라벨링 유형	내용요약(자연어)	라벨링 형식	JSON
데이터 활용 서비스	문서요약서비스, 주요문장추출서비스 등	데이터 구축년도/ 데이터 구축량	2021년/201,671

Train 데이터

- 종류

데이터 종류	원문 규모	어노테이션 규모	결과 규모		비고
			추출요약	생성요약	
뉴스기사	27,000	59,400	14,850	29,700	2~3문장 추출
			14,850		20% 추출
보도자료	20,000	44,000	11,000	22,000	2~3문장 추출
			11,000		20% 추출
역사_문화재	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
보고서	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
회의록	34,000	74,800	18,700	37,400	2~3문장 추출
			18,700		20% 추출
사설	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
간행물	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
연설문	40,000	88,000	22,000	44,000	2~3문장 추출
			22,000		20% 추출
문학	12,000	26,400	6,600	13,200	2~3문장 추출
			6,600		20% 추출
나레이션	10,000	22,742	5,640	11,371	2~3문장 추출
			5,731		20% 추출
총계	183,000	403,342	201,671	201,671	

※생성요약은 1문장 요약으로 100글자 이하로 구축

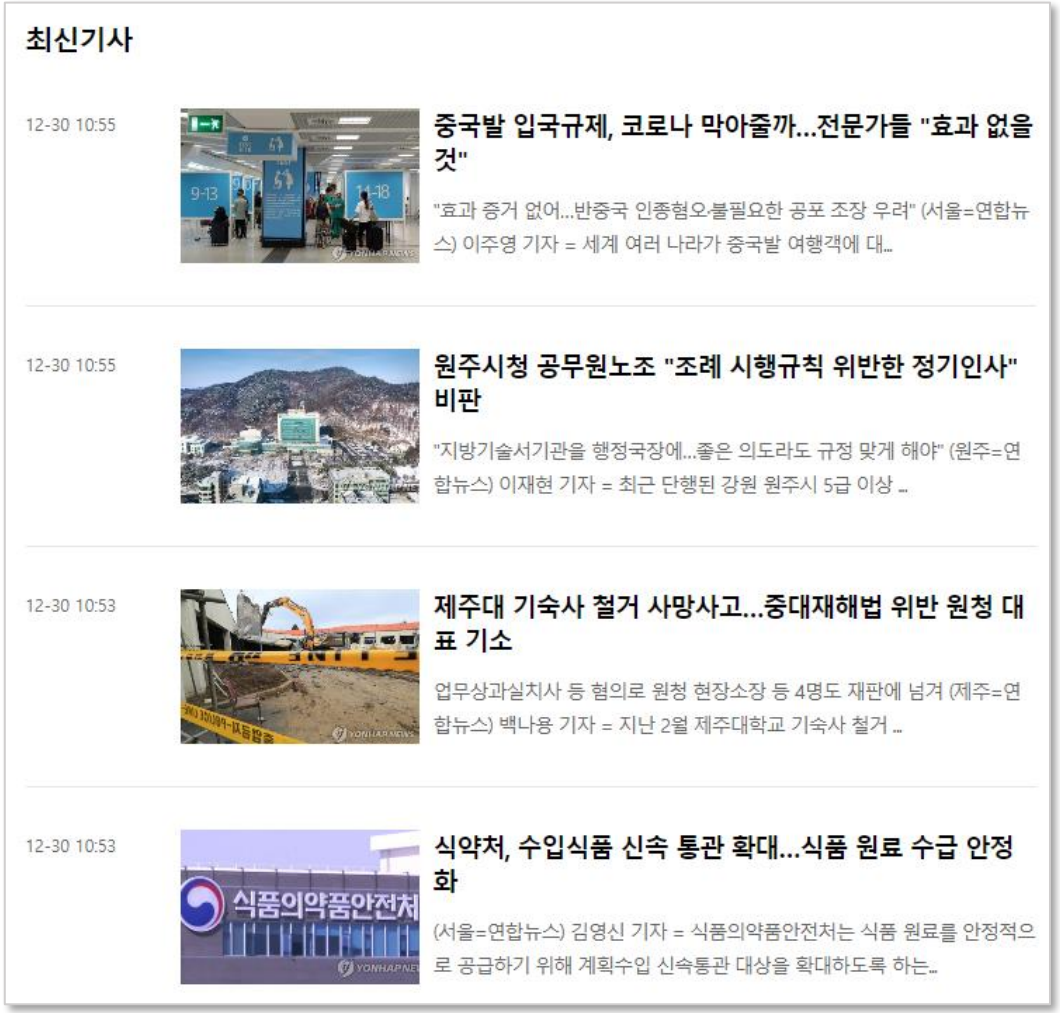
데이터

학습과정

성능개선

Test 데이터

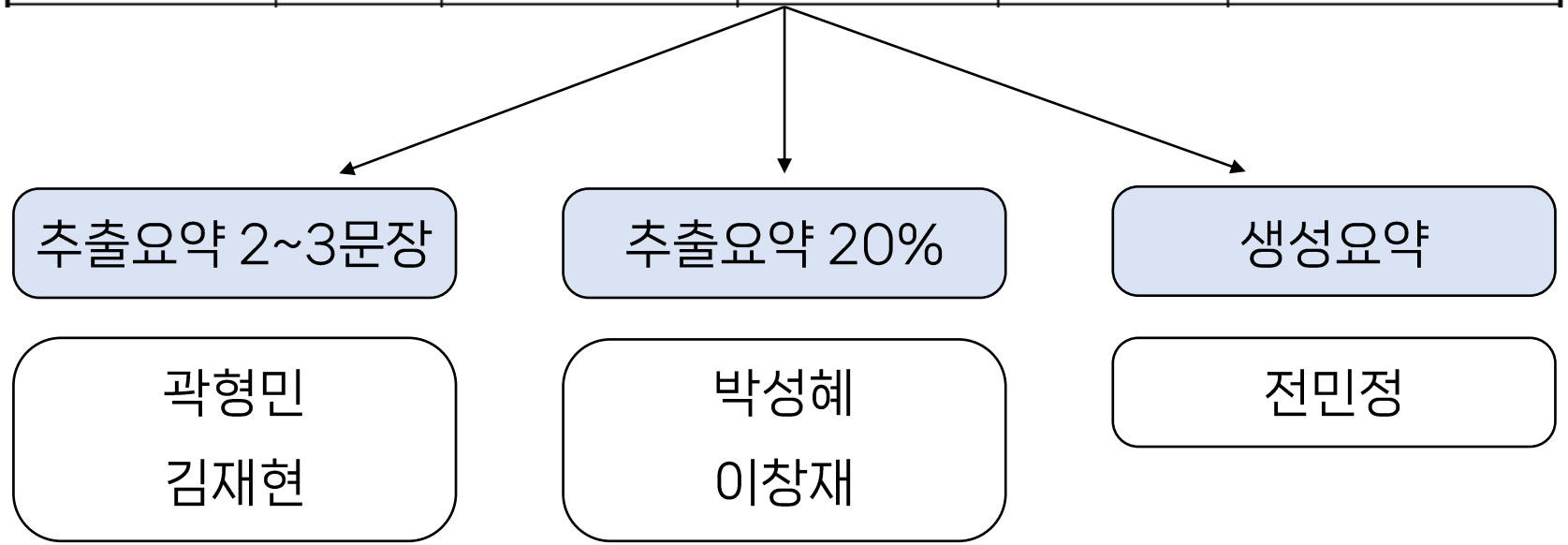
- 출처 : 연합뉴스
- 수집 방식 : 크롤링
- 규모 : 기사 10개



학습과정

데이터 분류

데이터 종류	원문 규모	어노테이션 규모	결과 규모		비고
			추출요약	생성요약	
뉴스기사	27,000	59,400	14,850	29,700	2~3문장 추출
			14,850		20% 추출



평가지표 1 ROUGE-L

- ROUGE-L 이란?
  → precision과 recall을 활용한 평가지표
  → text summarization, machine translation를 평가
  → 가장 긴 Sequence의 recall 을 구함
  → 만점 = 1

- 장점
  → 다양한 길이의 Sequence에서 stability와 reliability를 갖춤

- 단점
  1) 문맥 평가 불가
  2) 문장이나 표현이 일치하는지 여부만 파악 가능

평가지표 1 ROUGE-L

- 예시

정답문장 "**한화는 10 년 안에 우승 할 것이다.**"
생성문장:"**한화는 10 년 안에** 절대 **우승** 못 **할 것이다.** "

$$N_{정답문장} = 7$$

$$longest\_sequence = 한화는\ 10\ 년\ 안에\ 우승\ 할\ 것이다$$

$$N_{longest\_sequence} = 7$$

$$ROUGE - L = \frac{7}{7} = 1$$

- ROUGE-L에 대한 F1 score를 구함

$$2 \times \frac{ROUGE_L precision \times ROUGE_L recall}{ROUGE_L precision + ROUGE_L recall}$$

평가지표 2 Edit Distance

- Edit Distance 란?

  → 두 문자열 간의 유사도를 측정

  → 만점 = 0

	ø	c	a	k	e
ø	0	1	2	3	4
c	1	0	1	2	3
a	2	1	0	1	2
t	3	2	1	1	2

평가지표 3 BERT Score

- BERT Score 란?
  → BERT를 이용하여 두 문장의 의미를 비교
  → precision과 recall, F1 스코어를 활용한 평가지표
  → generation task에 유용하게 사용
  → 만점 = 1

- 장점
  1) 인간의 판단과 더 잘 연관됨
  2) 기존 메트릭보다 더 강력한 모델 선택 성능 제공

- 단점
  → BLEU 평가지표보다 긴 소요 시간

학습과정

모델 1 (TF)AutoModelForSeq2SeqLM

- 추출요약 데이터

Module	T5-small
Tokenizer	AutoTokenizer
Result	GPU 런타임 해제

- 생성요약 데이터

Module	digit82/kobart-summarization	csebuetnlp/mT5_multilingual_XLSum	Psyche/KoT5-summarization
Tokenizer	AutoTokenizer	AutoTokenizer	AutoTokenizer
Result	GPU 런타임 해제	ROUGE-L = 0.5	ROUGE-L = 1.0

학습과정

모델 2 (TF)BartForConditionalGeneration

- 추출요약 데이터

Module	gogamza/kobart-base-v1	kobart-summarization	noahkim/KoT5_news_summarization
Tokenizer	AutoTokenizer	PreTrainedTokenizer	BartTokenizer
Result	ROUGE-L = 0.4	ROUGE-L = 0.5	언더핏 로스가 높음

- 생성요약 데이터

Module	gogamza/kobart-base-v2
Tokenizer	BartTokenizer
Result	모델 자체 구현이 Sequential로 되어 있어서 실패

학습과정

모델 3 BERT

- 추출요약 데이터

Module	SKT/kobert-base-v1
Tokenizer	KoBERTTokenizer
Result	내부 모듈 호환 문제로 불가

학습과정

모델 4 (TF)T5ForConditionalGeneration

- 추출요약 데이터

Module	psyche/KoT5-summarization	KoT5-test	paust/pko-t5-small
Tokenizer	AutoTokenizer		T5TokenizerFast
Result	ROUGE-L = 0.75	ROUGE-L = 0.75 Edit_Distance = 0.74	ROUGE-L = 0.75

모델 1 BartForConditionalGeneration

- 추출요약 데이터

회차	1차	2차
Module	Kobart-summarization	
Tokenizer	PreTrainedTokenizerFast	
Data Scale	21,600	82,581
Result	ROUGE-L = 0.5 Edit_Distance = 2.85	ROUGE-L = 0.4 Edit_Distance = 1

모델 2 (TF)T5ForConditionalGeneration

- 추출요약 데이터

회차		1차	2차
Module		KoT5-test	
Tokenizer		AutoTokenizer	
Hyper parameter	Train batch	10	10
	Eval. batch	10	5
	Epoch	1	2
Result	ROUGE-L	0.75	**0.85**
	Edit Distance	0.72	**0.70**
	BERT Score	0.82	**0.83**

3. 결과 분석

원문

※ 연합뉴스 크롤링 데이터

노형욱 장관, 철도역 방역 · 운행 안전에 만전강조
14일 서울역 KTX 철도역사열차 방역실태 전반에 걸쳐 현장 점검\n 철도특별사법경찰대에는 방역수칙 위반자에 대한 무관용 원칙 대응 주문
노형욱 국토교통부 장관은 7월 14일 수도권 코로나19 방역의 관문이라고 할 수 있는 서울역을 방문하여, 철도역사 및 열차 방역실태 등을 점검하고, 관계자들을 격려하였다.
먼저, 노 장관은 한국철도공사로부터 수도권 거리두기 4단계 격상에 따른 전국 주요 역사 탑승 전 발열체크 등 강화된 철도분야 방역대책을 보고 받은 뒤, 코로나19 로 어려운 상황 속에서도 국민을 위해 힘써 주시는 노고에 감사드린다면서 관계자들을 격려하였다.
이어, KTX 방역현장을 둘러본 뒤, 지금까지 철도분야는 철두철미한 방역조치로 열차 내 감염사례가 단 한 건도 없었다라며, 강화된 방역대책이 승객에게 잘 안내되고, 이를 통해 실제 현장에서 잘 이행될 수 있도록 사명감을 갖고 총력을 기울여 줄 것을 강조하였다.
또한 방역으로 인해 자칫 운행 안전에 대한 경각심이 느슨해질 수 있다면서, 촘촘하고 치밀한 시설물 점검 및 차량의 정비를 통해 안전사고 예방에도 만전을 기해줄 것을 당부하였다.
이어, 서울지방철도특별사법경찰대를 방문한 노 장관은 철도시설 내 국민의 지킴이 역할과 해외입국자 KTX 전용칸 수송지원 업무를 차질 없이 수행하고 있는 철도경찰의 노고에 감사드린다면서, 코로나19로 인해 철도 승객 간 분쟁이 증가하는 등 철도경찰의 역할이 그 어느 때보다 중요한 때라며, 승객의 방역수칙 위반에 대해서는 무관용 원칙을 적용하여 국민들이 안심하고 철도를 이용할 수 있는 환경을 조성해 줄 것을 당부하였다

결과

노형욱 국토교통부 장관은 7월 14일 수도권 코로나19 방역의 관문이라고 할 수 있는 서울역을 방문하여, 철도역사 및 열차 방역실태 등을 점검하고, 관계자들을 격려하였다. KTX 방역현장을 둘러본 뒤, 지금까지 철도분야는 철두철미한 방역조치로 열차 내 감염사례가 단 한 건도 없었다라며, 강화된 방역대책이 승객에게 잘 안내되고, 이를 통해 실제 현장에서 잘 이행될 수 있도록 사명감을 갖고 총력을 기울여 줄 것을 강조하였다.

- 외부 요인

  1. GPU 부족
  2. 학습시간 부족
  3. 데이터 부족

- 내부 요인

  1. 길이가 긴 경우 소실되어 요약
  2. 모델 처리 속도 느림

- 개선방안

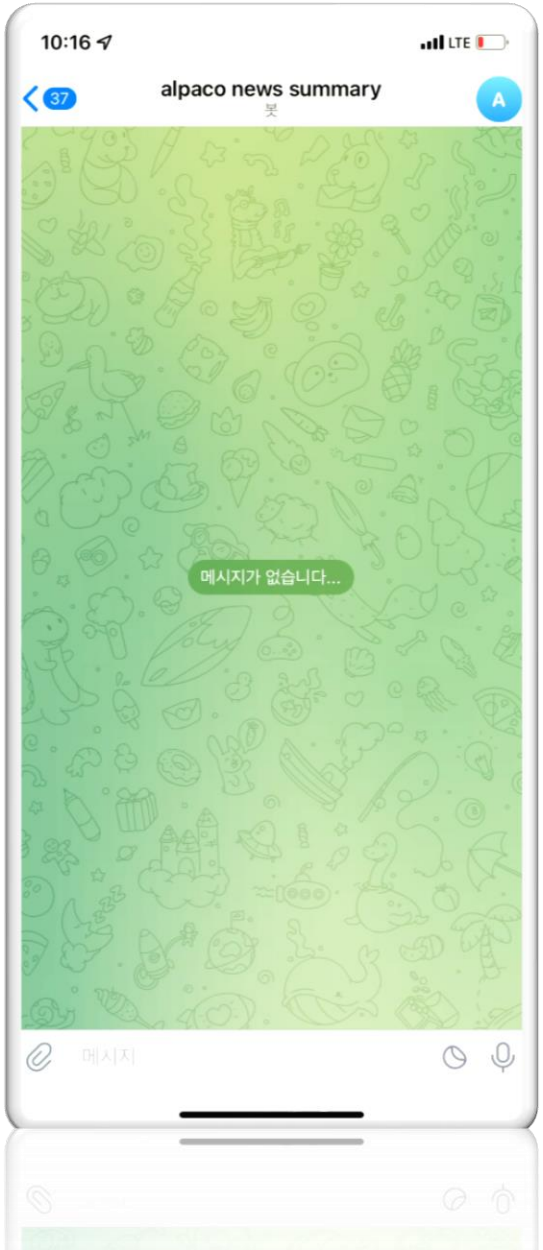
1. 짧은 글을 시작으로 긴 글을 읽을 수 있도록 원문 링크까지 첨부하여 접근성이 용이하도록 함

2. 연합뉴스를 기반으로 크롤링 하였으나 추후 여러 언론사를 추가하는 방향으로 확대

3. 최대 요약, 중간 요약 옵션을 지정하여 원하는 길이의 정보를 제공하도록 함

4. 실현 예시

요약봇 실현 예시



5. 기대 효과

1. 논문, 소설 등 뉴스 기사 외 요약이 필요한 분야에 적용 가능

2. 긴 글을 읽지 않는 성향이 높은 Z세대에게 글 읽는 연습 독려 가능

3. Z세대가 핵심 이슈에 쉽게 접근 가능

Thank you